Published in final edited form as:
Pac Symp Biocomput. 2007 ; : 269–280.

GeneRIF QUALITY ASSURANCE AS SUMMARY REVISION

ZHIYONG LU, **K. BRETONNEL COHEN**, and **LAWRENCE HUNTER**
Center for Computational Pharmacology, University of Colorado Health Sciences Center, Aurora, CO, 80045, USA

Abstract

Like the primary scientific literature, GeneRIFs exhibit both growth and obsolescence. NLM's control over the contents of the Entrez Gene database provides a mechanism for dealing with obsolete data: GeneRIFs are removed from the database when they are found to be of low quality. However, the rapid and extensive growth of Entrez Gene makes manual location of low-quality GeneRIFs problematic. This paper presents a system that takes advantage of the summary-like quality of GeneRIFs to detect low-quality GeneRIFs via a *summary revision* approach, achieving precision of 89% and recall of 77%. Aspects of the system have been adopted by NLM as a quality assurance mechanism.

1. Introduction

In April 2002, the National Library of Medicine (NLM) began an initiative to link published data to Entrez Gene entries via Gene References Into Function, or GeneRIFs. GeneRIFs consist of an Entrez Gene ID, a short text (under 255 characters), and the PubMed identifier (PMID) of the publication that provides evidence for the assertion in that text. The extent of NLM's commitment to this effort can be seen in the growth of the number of GeneRIFs currently found in Entrez Gene—there are 157,280 GeneRIFs assigned to 29,297 distinct genes (Entrez Gene entries) in 571 species as of June 2006. As we will demonstrate below, the need has arisen for a quality control mechanism for this important resource. GeneRIFs can be viewed as a type of low-compression, single-document, extractive, informative, topic-focussed summary [15]. This suggests the hypothesis that methods for improving the quality of summaries can be useful for improving the quality of GeneRIFs. In this work, we evaluate an approach to GeneRIF quality assurance based on a *revision model*, using three distinct methods. In one, we examined the recall of the system, using the set of all GeneRIFs that were withdrawn by the NLM indexers over a fixed period of time as a gold standard. In another, we performed a coarse assessment of the precision of the system by submitting system outputs to NLM. The third involved a fine-grained evaluation of precision by manual judging of 105 system outputs.

1.1. A fault model for GeneRIFs

Binder (1999) describes the *fault model*—an explicit hypothesis about potential sources of errors in a system [3]. Viewing GeneRIFs as summaries suggests a set of related potential sources of errors. This set includes all sources of error associated with extractive summarization (discussed in detail in [16]). It also includes deviations from the NLM's guidelines for GeneRIF production—both explicit (such as definitions of scope and intended content) and tacit (such as the presumed requirement that they not contain spelling errors).

E-mail: {Zhiyong.Lu, Kevin.Cohen, Larry.Hunter}@uchsc.edu.

Since the inception of the GeneRIF initiative, it has been clear that a quality control mechanism for GeneRIFs would be needed. One mechanism for implementing quality control has been via submitting individual suggestions for corrections or updates via a form on the Entrez Gene web site. As the size of the set of extant annotations has grown—today there are over 150,000 GeneRIFs—it has become clear that high-throughput, semi-automatable mechanisms will be needed, as well—over 300 GeneRIFs were withdrawn by NLM indexers just in the six months from June to December 2005, and data that we present below indicates that as many as 2,923 GeneRIFs currently in the collection are substandard.

GeneRIFs can be unsatisfactory for a variety of reasons:

- Being associated with a discontinued Entrez Gene entry
- Containing errors, whether minor—of spelling or punctuation—or major, i.e. with respect to content
- Being based only on computational data—the NLM indexing protocol dictates that GeneRIFs based solely on computational analyses are not in scope [7]
- Being redundant
- Not being informative—GeneRIFs should not merely indicate what a publication is about, but rather should communicate actual information
- Not being about gene function

This paper describes a system for detecting GeneRIFs with those characteristics. We begin with a corpus-based study of GeneRIFs for which we have third-party confirmation that they were substandard, based on their having been withdrawn by the NLM indexers. We then propose a variety of methods for detecting substandard GeneRIFs, and describe the results of an intrinsic evaluation of the methods against a gold standard, an internal evaluation by the system builders, and an external evaluation by the NLM staff.

In this work, we evaluate an approach to GeneRIF quality assurance based on a *summary revision model.* In summarization, *revision* is the process of changing a previously produced summary. [16] discusses several aspects of revision. As he points out (citing [5]), human summarizers perform a considerable amount of revision, addressing issues of semantic content (e.g., replacing pronouns with their antecedents) and of form (e.g., repairing punctuation). Revision is also an important component of automatic summarization systems, and in particular, of systems that produce extractive summaries, of which GeneRIFs are a clear example. (Extractive summaries are produced by “cutting-and-pasting” text from the original, and it has been repeatedly observed that most GeneRIFs are direct extracts from the title or abstract of a paper ([2,9,12,15]). This suggests using a “revision system” to detect GeneRIFs that should be withdrawn.

2. Related Work

GeneRIFs were first characterized and analyzed in [17]. They presented the number of GeneRIFs produced and species covered based on the LocusLink revision of February 13, 2003, and introduced the prototype GeneRIF Automated Alerts System (GRAAS) for alerting researchers about literature on gene products.

Summarization in general has attracted a considerable amount of attention from the biomedical language processing community. Most of this work has focussed specifically on medical text—see [1] for a comprehensive review. More recently, computational biologists have begun to develop summarization systems targeting the genomics and molecular biology domains [14,15]. GeneRIFs in particular have attracted considerable attention in the

biomedical natural language processing community. The secondary task of the TREC Genomics Track in 2003 was to reproduce GeneRIFs from MEDLINE records [9]. 24 groups participated in this shared task. More recently, [15] presented a system that can automatically suggest a sentence from a PubMed/MEDLINE abstract as a candidate GeneRIF by exploiting an Entrez Gene entry's Gene Ontology annotations, along with location features and cue words. The system can significantly increase the number of GeneRIF annotations in Entrez Gene, and it produces qualitatively more useful GeneRIFs than previous methods. In molecular biology, GeneRIFs have recently been incorporated into the MILANO microarray data analysis tool. The system builders evaluated MILANO with respect to its ability to analyze a large list of genes that were affected by overexpression of p53, and found that a number of benefits accrued specifically from the system's use of GeneRIFs rather than PubMed as its literature source, including a reduction in the number of irrelevant results and a dramatic reduction in search time [19]. The amount of attention that GeneRIFs are attracting from such diverse scientific communities, including not only bioscientists, but natural language processing specialists as well, underscores the importance of ensuring the quality of the GeneRIFs stored in Entrez Gene.

3. A corpus of withdrawn GeneRIFs

The remarkable increase in the total number of GeneRIFs each year (shown in Table 1) comes despite the fact that some GeneRIFs have been removed internally by the NLM. We compared the GeneRIF collection of June 2005 against that of December 2005 and found that a total of 319 GeneRIFs were withdrawn during that period. These withdrawn GeneRIFs are a valuable source of data for understanding the NLM's model of what makes a GeneRIF bad. Our analyses are based on the GeneRIF files downloaded from the NCBI ftp site[b] at three times over the course of a one-year period (June 2005, December 2005, and June 2006). The data and results discussed in this paper are available at a supplementary website[c].

3.1. Characteristics of the withdrawn GeneRIFs

We examined these withdrawn GeneRIFs, and determined that four reasons accounted for the withdrawal of most of them (see Figure 1).

1. **Attachment to a temporary identifier:** GeneRIFs can only be attached to existing Entrez Gene entries. Existing Entrez Gene entries have unique identifiers. New entries that are not yet integrated into the database are assigned a temporary identifier (the string *NEWENTRY*), and all annotations that are associated with them are provisional, including GeneRIFs. GeneRIFs associated with these temporary IDs are often withdrawn. Also, when the temporary identifier becomes obsolete, the GeneRIFs that were formerly attached to it are removed (and transferred to the new ID). 39% (123/319) of the withdrawn GeneRIFs were removed via one of these mechanisms.
2. **Based solely on computational analyses:** The NLM indexing protocol dictates that GeneRIFs based solely on computational analyses are not in scope. 37% (117/319) of the withdrawn GeneRIFs were removed because they came from articles whose results were based purely on computational methods (e.g., by prediction techniques) rather than traditional laboratory experiments.
3. **Typographic and spelling errors:** Typographic errors are not uncommon in the withdrawn GeneRIFs. They include misspellings and extraneous punctuation. 14%

[b] ftp://ftp.ncbi.nlm.nih.gov/gene
[c] http://compbio.uchsc.edu/Hunterlab/Zhiyong/psb2007

(46/319) of the withdrawn GeneRIFs contained errors of this type (41 misspellings and 5 punctuation errors).

4. **Miscellaneous errors:** 6% (20/319) of the withdrawn GeneRIFs were removed for other reasons. Some included the authors' names at the end, e.g., Cloning and expression of ZAK, a mixed lineage kinase-like protein containing a leucine-zipper and a sterile-alpha motif.***Liu TC, etc.*** Others were updated by adding new gene names or modifying existing ones. For example, the NLM replaced **POPC** with **POMC** in Mesothelioma cell were found to express mRNA for***[POPC]*** … for the gene POMC (GeneID: 5443).

5. **Unknown reasons:** we were unable to identify the cause of withdrawal for the remaining 4% (13/319) of the withdrawn GeneRIFs.

These findings suggest that it is possible to develop automated methods for detecting substandard GeneRIFs.

4. System and Method

We developed a system containing seven modules, each of which addresses either the error categories described in Section 3.1 or the content-based problems described in Section 1.1 (e.g. redundancy, or not being about gene function).

4.1. Finding discontinued GeneRIFs

Discontinued GeneRIFs are detected by examining the gene history file from the NCBI's ftp site, which includes information about GeneIDs that are no longer current, and then searching for GeneRIFs that are still associated with the discontinued GeneIDs.

4.2. Finding GeneRIFs with spelling errors

Spelling error detection has been extensively studied for General English (see [13]), as well as in biomedical text (e.g. [20]). It is especially challenging for applications like this one, since gene names have notoriously low coverage in many publicly available resources and exhibit considerable variability, both in text [10] and in databases [4,6]. In the work reported here, we utilized the Google spell-checking API[d]. Since Google allows ordinary users only 1,000 automated queries a day, it was not practical to use it to check all of the 4 million words in the current set of GeneRIFs. To reduce the size of the input set for the spell-checker, we used it only to check tokens that did not contain upper-case letters or punctuation (on the assumption that they are likely to be gene names or domain-specific terms) and that occurred five or fewer times in the current set of GeneRIFs (on the assumption that spelling errors are likely to be rare). (See Table 3 for the actual distributions of non-word spelling errors across unigram frequencies in the full June 2006 collection of GeneRIFs, which supports this assumption. We manually examined a small sample of these to ensure that they were actual errors.)

4.3. Finding GeneRIFs with punctuation errors

Examination of the 319 withdrawn GeneRIFs showed that punctuation errors most often appeared at the left and right edges of GeneRIFs, e.g. the extra parenthesis and period in). TNF-alpha promoter polymorphisms are associated with severe, but not less severe, silicosis in this population.(GeneID:7124) … or the terminal comma in Heart graft rejection biopsies have elevated FLIP mRNA expression levels, (GeneID:8837). We used regular expressions (listed on the supplementary web site) to detect punctuation errors.

[d] http://www.google.com/apis/

4.4. Finding GeneRIFs based solely on computational methods

Articles describing work that is based solely on computational methods commonly use words or phrases such as *in silico* or *bioinformatics* in their titles and/or abstracts. We searched explicitly for GeneRIFs based solely on computational methods by searching for those two keywords within the GeneRIFs themselves, as well as in the titles of the corresponding papers. GeneRIFs based solely on computational methods were incidentally also sometimes uncovered by the "one-to-many" heuristic (described below).

4.5. Finding similar GeneRIFs

We used two methods to discover GeneRIFs that were similar to other GeneRIFs associated with the same gene. The intuitions behind this are that similar GeneRIFs may be redundant, and that similar GeneRIFs may not be informative. The two methods involved finding GeneRIFs that are substrings of other GeneRIFs, and calculating Dice coefficients.

4.5.1. Finding substrings—We found GeneRIFs that are proper substrings of other GeneRIFs using Oracle.

4.5.2. Calculating Dice coefficients—We calculated Dice coefficients using the usual formula ([11]:202), and set our threshold for similarity at > 0.8.

4.6. Detecting one-to-many mappings

We used a simple hash table to detect one-to-many mappings of GeneRIF texts to publications (see category 6 in Table 2). We anticipated that this would address the detection of GeneRIF texts that were not informative. (It turned out to find more serious errors, as well—see the Discussion section.)

4.7. Length constraints

We tokenized all GeneRIFs on whitespace and noted all GeneRIFs that were three or fewer tokens in length. The intuition here is that very short GeneRIFs are more likely to be *indicative summaries*, which give the reader some indication of whether or not they might be interested in reading the corresponding document, but are not actually informative [16]—for example, the single-word text *Review*—and therefore are out of scope, per the NLM guidelines.

5. Results

5.1. Evaluating recall against the set of withdrawn GeneRIFs

To test our system, we first applied our system to the withdrawn GeneRIFs described in Section 3. GeneRIFs that are associated with temporary IDs are still in the curation process, so we did not attempt to deal with them, and they were excluded from the recall evaluation. To ensure a stringent evaluation with the remaining 196 withdrawn GeneRIFs, we included the ones in the miscellaneous and unknown categories. The system identified 151/196 of the withdrawn GeneRIFs, for a recall of **77%** as shown in Table 4. The system successfully identified 115/117 of the GeneRIFs that were based on solely computational results. It missed two because we limited our algorithm to searching only GeneRIFs and the corresponding titles, but the evidence for the computational status of those two is actually located in their abstracts. For the typographic error category, the system correctly identified 33/41 spelling errors and 3/6 punctuation errors. It missed several spelling errors because we did not check words containing upper-case letters. For example, it missed the misspellings *Muttant* (Mutant), *MMP-1o* (MMP-10), and *Frame-schift* (Frame-shift). It missed punctuation errors that were not at the edges of the GeneRIF, e.g. the missing space after the

semicolon in ***REVIEW:Association*** *of expression* … and the missing space after the comma in …***lymphocytes,suggesting*** *a role for trkB*…

5.2. 3rd-party evaluation of precision

The preceding experiment allowed us to evaluate the system's recall, but provided no assessment of precision. To do this, we applied the system to the entire June 2006 set of GeneRIFs. The system identified 2,923 of the 157,280 GeneRIFs in that data set as being bad. Table 2 shows the distribution of the suspicious GeneRIFs across the seven error categories. We then sent a sample of those GeneRIFs to NLM, along with an explanation of how the sample had been generated, and a request that they be manually evaluated. Rather than evaluate the individual submissions, NLM responded by internally adopting the error categories that we suggested and implementing a number of aspects of our system into their own quality control process, as well as using some of our specific examples to train the indexing staff regarding what is "in scope" for GeneRIFs (Donna Maglott, personal communication).

5.3. In-house evaluation of precision

We constructed a stratified sample of system outputs by selecting the first fifteen unique outputs from each category. Two authors then independently judged whether each output GeneRIF should, in fact, be revised. Our inter-judge agreement was 100%, suggesting that the error categories are consistently applicable. We applied the most stringent possible scoring by counting any GeneRIF that either judge thought was incorrectly rejected by the system as being a false positive. Table 5 gives the precision scores for each category.

6. Discussion and Conclusion

The kinds of revisions carried out by human summarizers cover a wide range of levels of linguistic depth, from correcting typographic and spelling errors ([16]:37, citing [5]) to addressing issues of coherence requiring sophisticated awareness of discourse structure, syntactic structure, and anaphora and ellipsis ([16]:78–81, citing [18]). Automatic summary revision systems that are far more linguistically ambitious than the methods that we describe here have certainly been built; the various methods and heuristics that are described in this paper may seem simplistic, and even trivial. However, a number of the GeneRIFs that the system discovered were erroneous in ways that were far more serious than might be suspected from the nature of the heuristic that uncovered them. For example, of the fifteen outputs in the stratified sample that were suggested by the one-to-many text-to-PMID measure (category 6 in Table 2), six turned out to be cases where the GeneRIF text did not reflect the contents of the article at all. The articles in question were relevant to the Entrez Gene entry itself, but the GeneRIF text corresponded to only one of the two articles' contents, presumably due to a cut-and-paste error on the part of the indexer (specifically, pasting the same text string twice). Similarly, as trivial as the "extra punctuation" measure might seem, in one of the fifteen cases the extra punctuation reflected a truncated gene symbol (*sir-2.1* became *−2.1*). This is a case of erroneous content, and not of an inconsequential typographic error. The word length constraint, simple as it is, uncovered a GeneRIF that consisted entirely of the URL of a web site offering Hmong language lessons —perhaps not as dangerous as an incorrect characterization of the contents of a PubMed-indexed paper, but quite possibly a symptom of an as-yet-unexploited potential for abuse of the Entrez Gene resource.

The precision of the length constraint was quite low. Preliminary error analysis suggests that it could be increased substantially by applying simple language models to differentiate GeneRIFs that are perfectly good indicative summaries, but poor informative summaries,

such as *REVIEW* or *3D model* (which were judged as true positives by the judges) from GeneRIFs that simply happen to be brief, but are still informative, such as *regulates cell cycle* or *interacts with SOCS-1* (both of which were judged as false positives by the judges).

Our assessment of the current set of GeneRIFs suggests that about 2,900 GeneRIFs are in need of retraction or revision. GeneRIFs exhibit the two of the four characteristics of the primary scientific literature described in [8]: growth, and obsolescence. (They directly address the problem of fragmentation, or spreading of information across many journals and articles, by aggregating data around a single Entrez Gene entry; linkage is the only characteristic of the primary literature that they do not exhibit.) Happily, NLM control over the contents of the Entrez Gene database provides a mechanism for dealing with obsolescence: GeneRIFs actually are removed from circulation when found to be of low quality. We propose here a data-driven model of GeneRIF errors, and describe several techniques, modelled as automation of a variety of tasks performed by human summarizers as part of the summary revision process, for finding erroneous GeneRIFs. Though we do not claim that it advances the boundaries of summarization research in any major way, it is notable that even these simple summary revision techniques are robust enough that they are now being employed by NLM: versions of the punctuation, “similar GeneRIF,” and length constraint (specifically, single words) have been added to the indexing workflow. Previous work on GeneRIFs has focussed on quantity—this paper is a step towards assessing, and improving, GeneRIF quality. NLM has implemented some of the aspects of our system, and has already corrected a number of the examples of substandard GeneRIFs that are cited here.

Acknowledgments

This work was supported by NIH grant R01-LM008111 (LH). We thank Donna Maglott and Alan R. Aronson for their discussions of, comments on, and support for this work, and the individual NLM indexers who responded to our change suggestions and emails. Lynne Fox provided helpful criticism. We also thank Anna Lindemann for proofreading the manuscript.

References

1. Afantenos S, Karkaletsis V, Stamatopoulos P. Summarization from medical documents: a survey. Artificial Intelligence in Medicine. Feb; 2005 33(2):157–77. Review. [PubMed: 15811783]
2. Bhalotia, G.; Nakov, PI.; Schwartz, AS.; Hearst, MA. Biotext report for the TREC 2003 genomics track. Proceedings of The Twelfth Text REtrieval Conference; 2003. p. 612
3. Binder, RV. Testing Object-Oriented Systems: Models, Patterns, and Tools. Addison-Wesley Professional; 1999.
4. Cohen, KB.; Dolbey, AE.; Acquaah-Mensah, GK.; Hunter, L. Contrast and variability in gene names; Proceedings of ACL Workshop on Natural Language Processing in the Biomedical Domain; Association for Computational Linguistics; p. 14-20.
5. Cremmins, ET. The Art of Abstracting. 2. Information Resources Press; 1996.
6. Fang, H.; Murphy, K.; Jin, Y.; Kim, JS.; White, PS. Proceedings of the BioNLP Workshop on Linking Natural Language Processing and Biology. Association for Computational Linguistics; Human gene name normalization using text matching with automatically extracted synonym dictionaries; p. 41-48.
7. GeneRIF: http://www.ncbi.nlm.nih.gov/projects/GeneRIF/GeneRIFhelp.html
8. Hersh, W. Information Retrieval: a Health and Biomedical Perspective. 2. Springer-Verlag; 2006.
9. Hersh, W.; Bhupatiraju, RT. TREC genomics track overview. Proceedings of The Twelfth Text REtrieval Conference; 2003. p. 14
10. Hirschman L, Colosimo M, Morgan A, Yeh A. Overview of BioCreative Task 1B: normalized gene lists. BMC Bioinformatics. 2005; 6(Suppl 1):S11. [PubMed: 15960823]
11. Jackson, P.; Moulinier, I. Natural Language Processing for Online Applications: Text Retrieval, Extraction, and Categorization. John Benjamins Publishing Co; 2002.

12. Jelier, B.; Schwartzuemie, M.; van der Fijk, C.; Weeber, M.; van Mulligen, E.; Schijvenaars, B. Searching for GeneRIFs: concept-based query expansion and Bayes classification; Proceedings of The Twelfth Text REtrieval Conference; 2003. p. 225

13. Jurafsky D, Martin JH. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition. Prentice Hall. Jan.2000

14. Ling, X.; Jiang, J.; He, X.; Mei, Q.; Zhai, C.; Schatz, B. Automatically generating gene summaries from biomedical literature; Proceedings of Pacific Symposium on Biocomputing; 2006. p. 40-51.

15. Lu Z, Cohen KB, Hunter L. Finding GeneRIFs via Gene Ontology annotations. Proceedings of Pacific Symposium on Biocomputing. 2006:52–63.

16. Mani, I. Automatic Summarization. John Benjamins Publishing Company; 2001.

17. Mitchell, JA.; Aronson, AR.; Mork, JG.; Folk, LC.; Humphrey, SM.; Ward, JM. Gene indexing: characterization and analysis of NLM's GeneRIFs. Proceedings of AMIA 2003 Symposium; 2003. p. 460-464.

18. Nanba, H.; Okumura, M. Producing more readable extracts by revising them. Proceedings of the 18th International Congress on Computational Linguistics (COLING-2000); p. 1071-1075.

19. Rubinstein R, Simon I. MILANO – custom annotation of microarray results using automatic literature searches. BMC Bioinformatics. 2005; 6:12. [PubMed: 15661078]

20. Ruch P, Baud R, Geissbuhler A. Using lexical disambiguation and named-entity recognition to improve spelling correction in the electronic patient record. Artificial Intelligence in Medicine. 2003; 29(2):169–84. [PubMed: 12957786]

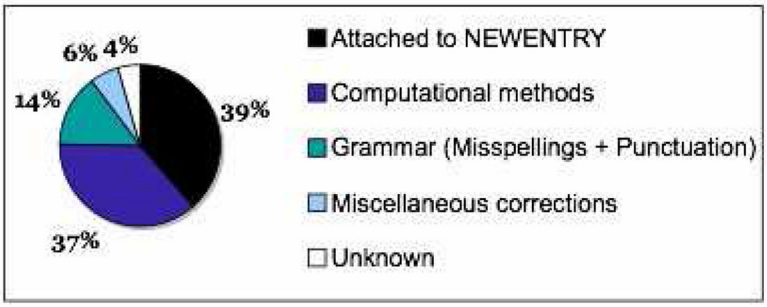

Figure 1.
Distribution of reasons for GeneRIF withdrawal from June to December 2005.

Table 1

GeneRIF statistics from 2000 to 2006. The second row shows the annual increase in new GeneRIFs. The third row shows the number of new species for the new GeneRIFs. The fourth row is the number of genes that gained GeneRIF assignments in the year listed in the first row. Note that although the gene indexing project was officially started by the NLM in 2002, the first set of GeneRIFs was created in 2000.

Year	**2000**	**2001**	**2002**	**2003**	**2004**	**2005**	**2006**[a]	**Sum**
New GeneRIFs	47	617	15,960	37,366	35,887	45,875	21,628	**157,280**
New Species	3	1	2	3	130	341	91	**571**
New Genes	34	529	6,061	6,832	5,113	7,769	2,959	**29,297**

[a]From January 2006 to June 2006

Table 2

A total of 2,923 suspicious GeneRIFs found in the June 2006 data. See Sections 4.5–7 for the explanations of categories 5–7.

No.	Category	GeneRIFs	GeneRIF example
1.	Discontinued	202	GeneID 6841: SVS1 seems to be found only in rodents and does not exist in humans
2.	Misspellings	1,754	GeneID 64919: CTIP2 mediates transcriptional repression with SIRT1 in **mammmalian** cells
3.	Punctuation	505	GeneID 7124: **).** TNF-alpha promoter polymorphisms are associated with severe, but not less severe, silicosis in this population.
4.	Computational results	19	GeneID 313129: characterization of rat Ankrd6 gene **in silico**; PMID 15657854: Identification and characterization of rat Ankrd6 gene **in silico**
5.	Similar GeneRIFs	209	GeneID 3937: two GeneRIFs for the same gene differ in the gene name in the parenthesis; Shb links SLP-76 and Vav with the CD3 complex in Jurkat T cells **(SLP-76)**
6.	One-to-many	67	A single GeneRIF text identification, cloning and expression is linked to two GeneIDs (217214 and 1484476) and two PMIDs (12049647, 15490124)
7.	Length Constraint	167	GeneID 3952: review; GeneID 135 molecular model; GeneID 81657: protein subunit function

Table 3

Distribution of non-word spelling errors across unigram counts.

Word Frequency	1	2	3	4	5
Spelling Errors	1,348	268	84	34	20

Table 4

Recall on the set of withdrawn GeneRIFs. Only the 196 non-temporary GeneRIFs were included in this experiment. Although we did not attempt to detect GeneRIFs that were withdrawn for miscellaneous or unknown reasons, we included them in the recall calculation.

Category	Total	True Positive	False Negative	Recall
Computational methods	117	115	2	98%
Misspellings	41	33	8	80%
Punctuation	5	3	2	60%
Miscellaneous	20	0	20	0
Unknown	13	0	13	0
Sum	**196**	**151**	**45**	**77%**

Table 5

Precision on the stratified sample. For each error category, a random list of 15 GeneRIFs were independently examined by the two judges.

No.	Category	True Positive	False Positive	Precision
1.	Discontinued	15	0	100%
2.	Misspellings	15	0	100%
3.	Punctuation	13	2	86.7%
4.	Computational methods	15	0	100%
5.	Similar GeneRIFs	15	0	100%
6.	One-to-many	15	0	100%
7.	Length constraint	5	10	33.3%
8.	**Overall**	**93**	**12**	**88.6%**